Journal of Applied Research in Memory and Cognition 9 (2020) 165–169

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Commentary

Why and how You Should Read Student Evaluations of Teaching

Nate Kornell*

Williams College, United States

If you think student evaluations of teaching (SET) tell you how much students learn from a teacher, get used to disappointment. SET scores are not correlated with learning (Uttl, White, & Gonzalez, 2017), they can create perverse incentives for teachers, and they are biased (Carpenter, Witherby, & Tauber, 2020). Criticisms like these have led some to suggest that SET should be abandoned. I argue that they should not. SET *are* biased, but most of this bias comes from the students, so there is no unbiased alternative. They *do* create bad incentives, but they create good incentives as well. Most important, SET measure student opinion. Knowing what the students think is not as good as knowing how good a teacher is, but student opinion does matter. SET are easy to misread, but useful if one knows how to use them. As Mike Tyson said, "It's good to know how to read, but it's dangerous to know how to read and not how to interpret what you're reading."[1]

Why should you listen to me? I am a professor of cognitive psychology at Williams College, a liberal arts college that prides itself on good teaching. I recently helped design a new SET form for Williams and I have written about SET before (Kornell & Hausman, 2016). Also, defending SET goes against my own bias: I hate SET. My first year at Williams College, I was shattered by my low ratings. Ten years later, my ratings are average, but like a rat that has been shocked repeatedly, I still freeze when I think about students judging me.

Three Shortcomings of SET

"The best argument against Democracy is a five-minute conversation with the average voter." –Winston Churchill[2]

Other articles have detailed the shortcomings of SET (e.g., Carpenter, Northern, Tauber, & Toftness, 2020; Neath, 1996; Pounder, 2007; Spooren, Brockx, & Mortelmans, 2013; Uttl et al., 2017). What follows is a condensed summary. The problem with SET is the students: They do not know how much they have learned, they are biased, and they (sometimes) prioritize good grades and entertainment over learning.

SET are Not Correlated with Learning

The most informative studies on SET and learning compare professors on a fair playing field by using the same exams for multiple courses and measuring learning in subsequent, related courses, because a student who learned more in Class 1 should do better in Class 2 (see Kornell & Hausman, 2016; Uttl et al., 2017). When teaching is measured in this way better teachers do not get better student ratings (Braga, Paccagnella, & Pellizzari, 2014; Carrell & West, 2010; Yunker & Yunker, 2003). This finding fits with an extensive literature on metacognition shows that people are not good at assessing their learning (Bjork, Dunlosky, & Kornell, 2013) especially with complex material (Dunlosky & Lipko, 2007; Carpenter, Wilford, Kornell, & Mullaney, 2013).

SET are Biased

In an article on how to get high SET ratings, Ian Neath started with "Tip 1: Be Male" (Neath, 1996). SET correlate with a teacher's gender, race, age, accent, and more (Carpenter, Northern et al., 2020). These biases can lead to unfair ratings and are a very serious problem.

SET do not cause these biases, however. The ratings are biased because the students are members of a biased society. Some argue that SET should be abandoned because they are biased. The problem is, what is the alternative? All methods of teacher evaluation require human judgment and there is no evidence that SET are more biased than other methods of teacher evaluation. Thus, there are only two options: use a biased system

Author Note

Nate Kornell, Department of Psychology, Williams College, United States.

This article was supported by James S. McDonnell Foundation Scholar Award 220020371. Hannah Hausman and Rachel Gerrard provided valuable comments on a draft of this article.

* Correspondence concerning this article should be addressed to Nate Kornell, Department of Psychology, Williams College, Williamstown, MA 01267, United States. Contact: nkornell@gmail.com.

to evaluate teaching or do not evaluate teaching at all. If one chooses to evaluate teaching, it is crucial to try to minimize bias and be aware of it when reading SET scores.[3]

SET Create Perverse Incentives

SET are not just measurements, they also influence teaching. Teachers often try to give the students what they want. The problem is, students do not always want what is best for their learning.

In psychology, students want "truth" that turns the world upside-down and leads to a simple take-home message that fits students' preexisting attitudes. Basically, they like TED talks. Unfortunately, these "truths" are rarely true. This problem is nicely captured when Dean Yeager criticizes parapsychologist Pete Venkman in Ghostbusters: "Doctor. . . Venkman. We believe that the purpose of science is to serve mankind. You, however, seem to regard science as some kind of dodge. . . or hustle. Your theories are the worst kind of popular tripe, your methods are sloppy, and your conclusions are highly questionable! You are a poor scientist, Dr. Venkman!" Dr. Venkman's defense is "But the kids love us!"[4]

Students also want good grades ("Tip 3: Grade Leniently"; Neath, 1996) and they like to be entertained ("Tip 13: Entertain" and "Tip 12: Gimmicks are Good"; Neath, 1996). Being entertaining does not always interfere with learning, but it can. For example, adding interesting but irrelevant tidbits of information to a text or lecture can decrease learning (Fries, DeCaro, & Ramirez, 2019; Harp & Maslich, 2005; Mayer, Griffith, Jurkowitz, & Rothman, 2008; Mayer, 2019). The bottom line is, SET can cause professors to give students what they want rather than what they need.

Three Reasons Why SET Should Not be Abandoned

"No one pretends that democracy is perfect or all-wise. Indeed, it has been said that democracy is the worst form of Government except all those other forms that have been tried from time to time." –Winston Churchill.

SET are, like democracy, the best way to do something that needs to be done. They have shortcomings, but they also have strengths: They measure student opinion, which is valuable; they create good incentives and prevent bad behavior; and there is no better alternative.

SET Create Positive Incentives and Prevent Bad Behavior

There are things that students want that are also good for learning. Some of these things go unnoticed unless they are missing. For example, students like it when teachers are organized ("Tip 2: Be Organized"; Neath, 1996), show up for class, hold office hours, know the material, and answer questions carefully. They want fair tests, impartial grading, reading assignments that are interesting and educational, and so on. Accomplishing these things requires time and effort on the part of the professor. SET rewarding professors for doing these things, and as such they create positive incentives that increase learning.

Given that SET create good and bad incentives, the question becomes which are stronger (Skinner, 1953). In other words, if SET were abandoned, would the quality of teaching get better or worse? I believe the worst teaching would get even worse.

The scariest professor is the one who does not care about student opinion. For example (these are real examples) the professor who, during office hours, slowly and cruelly tears a student's first paper to shreds; who mocks students in class; who gives bad grades to papers that challenge her personal religious beliefs; who states "facts" that are nowhere near correct; who never answers email and does not grade any of the homework until after finals; or the professor who assigns a book that he wrote and self-published, even though it is terribly out of date, because he makes money when students buy it.

If SET were abandoned, these things would happen more often because professors would have less reason to care about their students' opinions. Of course, if SET are retained we face the opposite problem, professors who use chocolate and funny irrelevant stories to buy students' affection. Although neither is ideal, the professor who cares too much seems preferable to the professor who cares too little. In other words, SET are a form of quality control. Like Olympic drug tests, they might be holding back a flood of bad behavior.

In summary, SET create good and bad incentives. I argue that SET might be doing more good than harm, mostly because it is easier to be a bad teacher than a good one and SET incentivize putting effort into one's courses.

SET Measure Student Opinion

If a professor is like a chef, then SET tell you how much diners like the food. This is less important than how healthy the food is (i.e., how much students learn), but it still matters. Giving students a good experience is a valid goal. If the diners hate the food—even if it is healthy—don't you want to try to do better? And don't you want to know?

It is important to recognize that student opinion is not the same as teaching quality. But if we did not have this information, many of us (including me) would be asking ourselves, "I wonder what the students think?" To stop measuring student opinion is a way of telling students that you neither respect their opinion nor care whether they are happy with their classes.

One might say learning is what matters, not happiness. I think colleges should aspire to both. The mission of a college should be about more than learning in class. Professors have the ability to motivate students, to inspire them, and to play all sorts of

---

[1] https://www.brainyquote.com/quotes/mike_tyson_369903

[2] Churchill never actually said this according to https://winstonchurchill.org/publications/finest-hour/finest-hour-141/red-herrings-famous-quotes-churchill-never-said/.

[3] Other sources of bias are remedied more easily. For example, instructors get higher ratings if they are present at their evaluation ("Tip 4: Be Present at Your Evaluation," Neath, 1996), if they give SET instructions in a charming, humorous way ("Tip 6: Provide the 'Correct' Instructions"), and avoid giving surveys after a test ("Tip 5: Administer Ratings Before Tests"). They also get higher ratings if they give out candy (Youmans & Jee, 2007). These tricks shift the basis of the ratings away from what actually matters in a course, which is one reason why professors should not administer their own SET forms in class.

[4] https://www.imdb.com/title/tt0087332/quotes/qt0475985

other roles. They are role models, for better or worse. One reason to use SET is to identify professors who—by virtue of getting very low SET ratings—are not good role models and do not create a positive student experience. This is all the more important because of the power imbalance between a student and a professor. An unkind word from a professor can be devastating.

If a teacher's students are learning, one must ask, does it matter what the students think of the professor? If it does, then SET are useful because they provide this information. If student opinion does not matter, then SET have little value. I believe that learning is the most important thing, but assuming that learning is happening, I also believe that hiring and promoting professors who get high ratings creates a faculty whose courses students want to take—and that is the kind of faculty I hope my children encounter in college.

There is No Better Alternative

There are four main ways to evaluate teaching quality: numerical SET, qualitative student interviews or surveys, examining course materials, and peer evaluations by professors. Each method has pros and cons and the best way to evaluate teaching is to use all four. Next I discuss the strengths and weaknesses of qualitative surveys and peer evaluations.

Some propose that qualitative surveys should replace numerical SET because the latter are flawed. Unfortunately, the numerical ratings are not the problem; if the students' opinions are inaccurate and biased, asking questions a different way will not fix the problem. If anything, non-numerical surveys are worse. A biased (e.g., sexist) reader given a series of paragraphs written by students has more freedom to misconstrue, cherry-pick, and reinterpret than if they are given numerical data. Even for an unbiased reader, qualitative ratings are tricky; two comments out of 50 are especially vivid and stand out, one might end up being influenced more by those two comments than the other 48. When you compute a mean, it weights every student equally.

Qualitative ratings also have advantages. The two biggest are that they can help explain why low numerical scores are low (or high ones are high) and that they give feedback to the teacher. The best solution is probably to use qualitative surveys to serve these functions, and numerical SET to capture student opinion accurately in aggregate.

There is also a third alternative: Peer evaluations of teaching. These also have major advantages: Professors know the material and they are (relatively) expert teachers. Peer evaluations also have disadvantages, however. Peers typically show up to just 1-3 classes instead of a whole semester's worth. College professors don't necessarily know more than students about which strategies are good for learning (Morehead, Rhodes, & DeLozier, 2016). And even if the peer sits in on classes and looks carefully at the course syllabus and materials (the fourth alternative, which is sometimes done by the peer observer), they miss a lot of class experiences that affect learning, such as visiting office hours, being graded, receiving feedback, taking exams, and doing reading assignments and homework (Lang, 2019).

Peer evaluations also suffer because peers pull their punches. Professors often have to live down the hallway from the person they are evaluating, so it is in the observer's self-interest to be kind—even when it is not appropriate. Peer bias is also an issue. Professors are not immune to societal bias. They also might have pre-existing feelings about the person they are observing. These feelings can be strong (you might really like or really dislike your colleague) and they are likely to influence peer evaluations because of halo effects (Nisbett & Wilson, 1977), confirmation bias (Nickerson, 1998), and motivated reasoning (Kunda, 1990). Structured peer observation rubrics are an alternative, but peer observers can use rubrics inconsistently, even after intensive training and practice (Amrein-Beardsley & Popp, 2012).

Peer evaluations, in short, can be a good supplement to SET. The upside is that they are written by experts; the downside is that they are often overly positive and can be biased. But they are not a replacement for SET because a professor cannot tell you student opinion.

In summary, numerical SET are good for identifying unpopular teachers, qualitative student surveys help explain why ratings came out how they did, and peer evaluations of teaching and course materials provide a perspective that is not based on student opinion. The best way to evaluate teaching is to consider all of these methods together and keep their weaknesses in mind.

How to Read SET

SET have serious weaknesses. Interpret them carefully. Here are two tips on how to do that: Look for big problems and know the audience.

Use SET to Find Problems

William Henry Seward said "Public opinion, in every country, is a capricious sea. Whoever attempts to navigate it is liable to be tossed about by storms" (Seaward, 1873). This is true of SET, where turbulence is to be expected. Small differences (e.g., between 3.9 and 4.1) are generally just random fluctuations and should not be treated as meaningful. Professors whose average ratings are in the middle of the rating distribution—even if they are a little below average—are usually all about the same.

In my opinion, most professors should be placed into two basic categories based on SET: worryingly low or fine. Unfortunately, a professor who is too fluent and tries too hard to please will be classified as "fine." This concern is absolutely real and SET is not the answer to this problem. But SET still have value if they identify professors whose popularity with students is markedly and consistently low. This is important for two reasons: Tenure decisions can turn on whether a professor's scores are too low (but not too high) and low ratings usually are a sign that whatever is happening is worse than just a professor making students learn a lot.

Know the Audience

If a professor gets good ratings, it does not necessarily mean she taught well. It means she gave the students what they wanted. These are different things. When reading another professor's

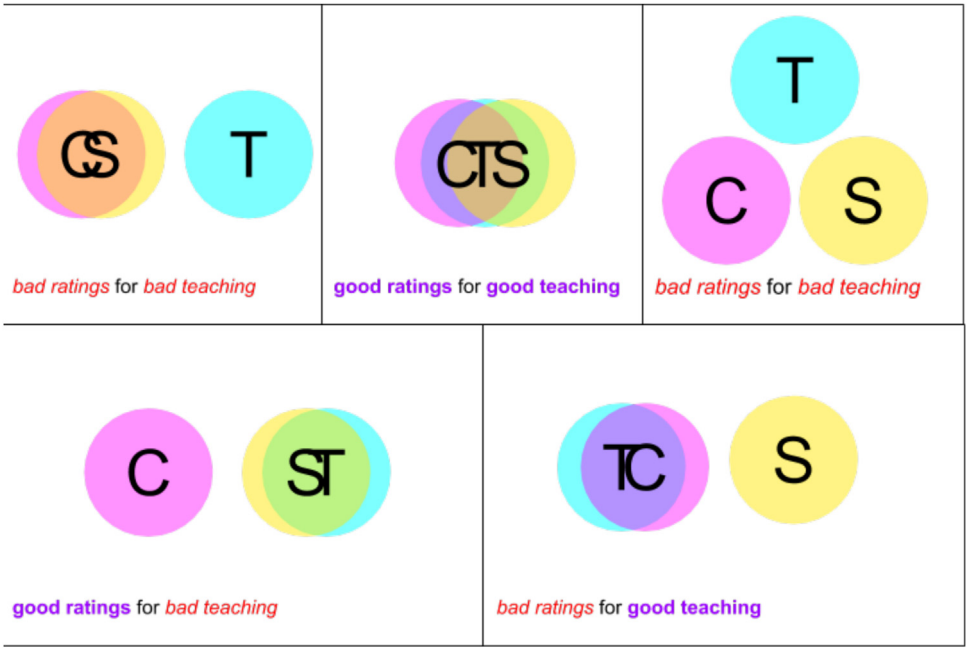

Figure 1. A teacher's goals (T) control what happens in a course. When they overlap with the college's goals (C) the result is good teaching. When they overlap with the students' goals (S), the result is high ratings. SET can be accurate (top row) or inaccurate (bottom row) depending on the pattern of overlap. As examples, in the top-left panel the students and college want a course about practical statistics but the teacher mostly talks about statistical theory. Top-middle: the course is about practical statistics and everyone is happy. Top-right: the students want a course about practical statistics, the college wants a class about statistical theory, and the teacher covers research design. Bottom-left: The college wants a serious course on statistics but the teacher and students want easy As in a course on statistical fluff. Bottom-right: The teacher and college want a serious course on statistics but the students want easy As in a course on statistical fluff.

SET scores, it helps to ask oneself what the students wanted and recognize that high scores probably mean they got it.

Different people want different things. When students want to learn, then high ratings are likely to be correlated with high amounts of learning. When they want easy homework, high ratings will tend to be correlated with easy homework. Figure 1 displays situations where SET are likely to be more or less meaningful depending on the overlap between the goals of the student and the institution.

Ask yourself what students at your institution want. A professor who gets high SET scores (i.e., is popular with students) at a party school might not be achieving the most learning. SET ratings might have more meaning at a school where students want to work hard and learn a lot. The reality is that both kinds of school exist and the proper way to interpret SET may vary accordingly.

Also ask about the course. The first question is usually, is the course required? If it is, there is a chance what the students wanted was to avoid having to take it ("Tip 11: Required Classes are Bad"; Neath, 1996). Students are more likely to enjoy a course—or a movie or a meal—if they chose it voluntarily. This is not the professor's fault and she should not be penalized for it. A professor who teaches two courses equally well should be expected to get lower ratings in a required course.

Students also like small, upper-level courses within their major ("Tip 8: The Smaller the Better"; "Tip 9: The Higher [course number] the Better"; "Tip 10: Cross-Listings are Bad"; Neath, 1996) Some courses within a major are not what the students want, though, such as required statistics courses in the psychology major. Students also want you to agree with them ("Tip 16: Be Like Your Students"). Ratings might be undeservedly low if a professor questions the orthodoxy of the students, or undeservedly high if she kowtows to it.

Summary

Reading SET requires a mindset adjustment. They tell you student opinion, and although this information is valuable, it is not the same as how good a teacher is. SET should not be overinterpreted; they should be used to identify teachers who are consistently and clearly unpopular with students. SET should be combined with other forms of evaluation, such as qualitative student surveys, peer observations, and teaching portfolios, which can all help pinpoint what is going right and wrong.

Author statement

Nate Kornell was responslible for the conception and writing of this article. The author reports no conflicts of interest.

Keywords: Student evaluations of teaching, Opinion, Bias, Learning, Tenure, Judgment

References

Amrein-Beardsley, A., & Popp, S. E. O. (2012). Peer observations among faculty in a college of education: investigating the summative and formative uses of the Reformed Teaching Observation Protocol (RTOP). *Educational Assessment, Evaluation and Accountability*, *24*, 5–24. http://dx.doi.org/10.1007/s11092-011-9135-1

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: beliefs, techniques, and illusions. *Annual Review of Psychology*, *64*, 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review*, *41*, 71–88. http://dx.doi.org/10.1016/j.econedurev.2014.04.002

Carpenter, S. K., Northern, P. E., Tauber, S. U., & Toftness, A. R. (2020). Effects of lecture fluency and instructor experience on students' judgments of learning, test scores, and evaluations of instructors. *Journal of Experimental Psychology: Applied*, *26*, 26–39. http://dx.doi.org/10.1037/xap0000234

Carpenter, S. K., Wilford, M. M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review*, *20*(6), 1350–1356. http://dx.doi.org/10.3758/s13423-013-0442-z

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On students' (mis)judgments of learning and teaching effectiveness. *Journal of Applied Research in Memory and Cognition*, *9*, 137–151.

Carrell, S. E., & West, J. E. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *Journal of Political Economy*, *118*, 409–432.

Dunlosky, J., & Lipko, A. (2007). Metacomprehension: a brief history and how to improve its accuracy. *Current Directions in Psychological Science*, *16*, 228–232. http://dx.doi.org/10.1111/j.1467-8721.2007.00509.x

Fries, L., DeCaro, M. S., & Ramirez, G. (2019). The lure of seductive details during lecture learning. *Journal of Educational Psychology*, *111*, 736–749. http://dx.doi.org/10.1037/edu0000301

Harp, S. F., & Maslich, A. A. (2005). The consequences of including seductive details during lecture. *Teaching of Psychology*, *32*, 100–103. http://dx.doi.org/10.1207/s15328023top3202_4

Kornell, N., & Hausman, H. (2016). Do the best teachers get the best ratings? *Frontiers in Psychology*, *7* http://dx.doi.org/10.3389/fpsyg.2016.00570

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498. http://dx.doi.org/10.1037/0033-2909.108.3.480

Lang, J. M. (2019). We don't trust course evaluations, but are peer observations of teaching much better? *The Chronicle of Higher Education*, https://www.chronicle.com/article/We-Don-t-Trust-Course/246432

Mayer, R. E. (2019). Taking a new look at seductive details. *Applied Cognitive Psychology*, *33*, 139–141. http://dx.doi.org/10.1002/acp.3503

Mayer, R. E., Griffith, E., Jurkowitz, I. T. N., & Rothman, D. (2008). Increased interestingness of extraneous details in a multimedia science presentation leads to decreased learning. *Journal of Experimental Psychology: Applied*, *14*, 329–339. http://dx.doi.org/10.1037/a0013835

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory*, *24*, 257–271. http://dx.doi.org/10.1080/09658211.2014.1001992

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports*, *78*, 1363–1372. http://dx.doi.org/10.2466/pr0.1996.78.3c.1363

Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. http://dx.doi.org/10.1037/1089-2680.2.2.175

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*, 250–256. http://dx.doi.org/10.1037/0022-3514.35.4.250

Pounder, J. S. (2007). Is student evaluation of teaching worthwhile? *Quality Assurance in Education*, *15*(2), 178–191. http://dx.doi.org/10.1108/09684880710748938

Seaward, W. H. (1873). *William H. Seward's travels around the world.* New York: D. Appleton.

Skinner, B. F. (1953). *Science and human behavior*. New York: Macmillan.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research*, *83* http://dx.doi.org/10.3102/0034654313496870

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation*, *54*, 22–42. http://dx.doi.org/10.1016/j.stueduc.2016.08.007

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology*, *34*, 245–247. http://dx.doi.org/10.1080/00986280701700318

Yunker, P. J., & Yunker, J. A. (2003). Are student evaluations of teaching valid? Evidence from an analytical business core course. *Journal of Education for Business*, *78*, 313–317. http://dx.doi.org/10.1080/08832320309598619

Received 21 February 2020;
received in revised form 25 February 2020;
accepted 25 February 2020
Available online xxx